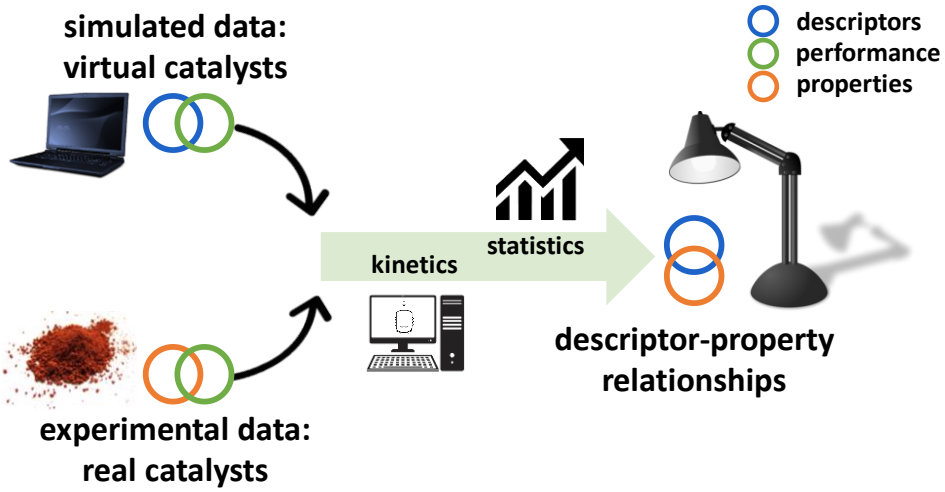
Descriptor - property relationships in heterogeneous catalysis: exploiting synergies between statistics and fundamental kinetic modelling

Laura Pirro[†], *Pedro S. F. Mendes*[†], *Stijn Paret*[†], *Bart D. Vandegehuchte*[‡], *Guy B. Marin*[†], *Joris W. Thybaut*[†,*]

† Laboratory for Chemical Technology, Ghent University, Technologiepark 125, B-9052 Ghent, Belgium

‡ Total Research & Technology Feluy, Zone Industrielle Feluy C B-7181, Seneffe, Belgium

*joris.thybaut@ugent.be



Combined kinetic and statistical approach to shed light on the link between kinetically-relevant descriptors and easily tuneable catalyst properties.

Abstract

Kinetics-driven design of heterogeneous catalysts is a promising methodology to optimize material performances for a reaction of interest. However, most of the fundamental kinetic models suffer from a limited applicability due to ill-defined relationships between catalyst features which impact the kinetics, i.e. catalyst descriptors, and properties determined from material characterization. In order to overcome this limitation, a comprehensive methodology is proposed, combining kinetic simulations with a selection of statistical tools. The aim is to identify similarities between experimental data and simulated performances, and to assess the significance of the descriptor-property relationships that can be established. Oxidative Coupling of Methane (OCM) was selected as case study to demonstrate this methodology. The obtained qualitative relationships indicated that the electronic properties of surface oxygen species are key in the optimization of OCM catalysts. This proof of concept highlights the proposed methodology as a tool for catalyst design for a broad variety of reactions, provided that a kinetic model and (a limited amount of) experimental data are available.

1. Introduction

Catalysts are the enablers of over 90% of chemical processes around the globe[1] and progress in the field of heterogeneous catalysis goes historically hand in hand with the evolution of human kind[2]. It is not surprising that the quest for new catalytic materials represents a continuously ongoing research challenge[3]. Preferably, a link is established between the structure of solid materials and their performance in a desired reaction, with the aim of speeding up catalyst optimization. The design of a catalyst is quite challenging, given the complexity of solid materials in general and of heterogeneous catalysts in particular, and the requirement of extensive knowledge in several scientific domains, ranging from surface science to chemical reaction engineering[2].

Several strategies have been pursued so far[4], including high-throughput experimental screening[5] and computational chemistry[6]. These approaches are more and more often complemented by machine learning techniques[7]. At the same time, the exploitation of experimental data present in literature[8-10] is growing in the field of materials discovery, to extract as much knowledge as possible from already available information[11, 12]. Notable examples of data mining in the domain of heterogeneous catalysis are provided by Rothenberg and coworkers[13-15], who set the scene for the machine learning applications that followed[16]. Recently, a comprehensive overview of the efforts which were devoted to the computer-assisted task of extracting knowledge from catalytic data, defined as catalysis informatics[12], was provided by Medford et al.[17].

Among the various approaches, those relying on mechanistic understanding remain most interesting in the development of new catalytic materials[18-20]. In fact, by incorporating the impact of process conditions and catalyst properties on reaction pathways in a chemically sound manner, kinetics-driven catalyst design[21, 22] can be extended to broader ranges of process conditions and catalytic materials, thus providing a competitive advantage over purely empirical and statistical methods. The leading role of kinetics in the optimization strategies for heterogeneous catalysts is testified also by the existence of a dedicated software for a descriptor-based microkinetic analysis of catalytic trends, CatMAP[23].

According to the terminology introduced by Katare et al.[24] in kinetics-driven catalyst design, and adopted more recently by Takahashi et al.[12] for data-to-design strategies, the most difficult task is to move from a *forward* problem, i.e. simulation of catalytic performances via kinetic modelling, to an *inverse* problem, namely the design of new materials based on model predictions. In previous work from our group[25], this is described as the closing step of the catalyst design cycle. In fact, the search for new catalytic materials can be seen as an iterative process, based on four steps: first, a limited library of catalysts is synthesized; subsequently,

these catalysts are tested for the reaction of interest; the third step consists of kinetic model regression by estimating so-called 'kinetic' and 'catalyst' descriptors. The former are parameters relevant to the kinetics of the reaction irrespective of the catalyst, while the latter are kinetically significant catalyst features that specifically account for the impact of the catalytic material on the reaction kinetics[26]. In the fourth step, catalytic performances are optimized via the developed kinetic model, and a set of catalyst descriptors corresponding to these optimized performances is identified. The next iteration starts via the synthesis of a new generation of catalysts exhibiting descriptors closer to optimal performance. The closure of this catalyst design cycle requires translating the (optimized set of) catalyst descriptors to properties which can be determined from material characterization and easily tuned during the synthesis, and this remains an intricate task.

Some successful examples of kinetics-driven design can be found in the field of hydroconversion[27-30]. However, several cases are reported in literature in which a smooth transition from descriptors to properties could not be achieved. For example, some catalyst descriptors were introduced and subsequently optimized via fundamental kinetic modelling for the respective reaction (propane aromatization[24], ethylene glycol decomposition[31], oxidative coupling of methane[32], ethylene oligomerization[25]). Unfortunately, guidelines for synthesis recipes or for actual properties of these better-performing materials could not be provided. One of the major risks of this approach is optimizing performances via unrealistic sets of descriptors, corresponding to materials which cannot be realized[33]. A potential solution to relate descriptors and structural properties of actual catalytic materials is offered by density functional theory[34]. While the approach based on volcano curve construction has been proven useful, for instance, in identifying alternative transition metal surfaces for acetylene selective hydrogenation[35] and CO_2 reduction[36], the present computational power does not allow including the whole complexity of a heterogeneous catalyst and/or mechanisms at once. The issue of computational

efficiency has been partially addressed, in an application to methane steam reforming catalysts[37], by applying the Degree of Rate Control method[38]. This approach, which focuses the screening efforts in the direction of a small number of rate determining species, has proved indeed to be more efficient. However, limitations still arise when the properties of the catalytic materials which are screened are too different from the reference materials for which the microkinetic model is originally constructed.

The methodology proposed in the present work aims at establishing the missing descriptor-property link in kinetics-driven catalyst design, by combining fundamental kinetic modelling with carefully chosen tools from data mining and statistical analysis. The methodology is applied to literature data on of the reactions for which the search for active, selective and stable catalysts is still an open challenge: the Oxidative Coupling of Methane (OCM)[39]. OCM aims at oxidatively converting methane, typically from natural gas, into C_2 hydrocarbons, preferably ethylene[40]. The process comprises a complex network of gas-phase and surface catalytic reactions[41], and suffers from low yields due to partial and total oxidation reactions of both reactant and C_2 products[42]. There are two main reasons why OCM has been chosen as case study in the present work. I) A vast number of performance data[43] on a broad database of catalysts has been acquired over 35 years of research, e.g. via high-throughput testing[44 45]. Some authors tried to identify trends in these datasets which could suggest favourable catalyst compositions. Statistical and machine learning techniques[46-48] were used and were, more recently, combined with rules guided by chemical knowledge[49]. As previously mentioned, these techniques rarely incorporate a mechanistic understanding of the reaction, and the impact of process conditions is included only from a statistical perspective, thus hampering reliable extrapolations to industrially relevant operating ranges. These considerations, in addition to the absence of a real breakthrough in catalyst design for OCM in the recent years[43], suggests that new methodologies are still needed to successfully address this challenge. II) A fundamental

kinetic model was developed in our research group for this reaction[50], allowing to reproduce experimental data acquired over five different catalysts[51], and this constitutes a vital pre-requisite for the methodology herein proposed.

A general overview of the methodology and the specific tools are described in section 2. In section 3, three OCM datasets are investigated. Section 4 includes a discussion about the descriptor – property relationships that could be extracted for the OCM case study and the lessons learnt from this proof of concept. Some conclusions are drawn in section 5 and the perspectives of the elaborated methodology are discussed.

2. Methodology development

The present methodology relies on the concepts of *real* and *virtual* catalysts, which are represented by orange and blue dashed areas, respectively, in Figure 1. *Real* catalysts are materials with reported chemical composition and possibly other measurable and tuneable properties (orange circle), that have been experimentally tested at a certain set of process conditions. Hence, their performance (green circle) in the reaction of interest is available. The number of *real* catalysts that are included in the analysis is indicated by q. *Virtual* catalysts are computer or '*in silico*' representations of catalytic materials: each *virtual* catalyst is represented by a vector of m catalyst descriptors D_j (blue circle), with j going from 1 to m. The descriptors are key parameters to simulate the performance (green circle) of the *virtual* catalysts via a fundamental kinetic model. In the present work, the number of *virtual* catalysts that are included in the analysis is indicated by n. n *virtual* catalysts, hence, constitute an *in silico* library of catalysts, which lies at the heart of the present methodology. Libraries of descriptors for materials are not uncommon, such as in the work of Farrusseng and co-workers[52, 53]. Their studies on Quantitative Structure/Properties Activity Relationships (QSAR/QPAR) for heterogeneous catalysts, and in particular the concept of *discovery* vs *targeted* libraries, have

been a source of inspiration for the present work. A *discovery* library is meant to be as diverse as possible, i.e. the *virtual* catalysts therein reproduce a broad range of reaction performances. A *targeted* library, on the other hand, contains *virtual* catalysts that perform similarly as a designated group of *real* catalysts.

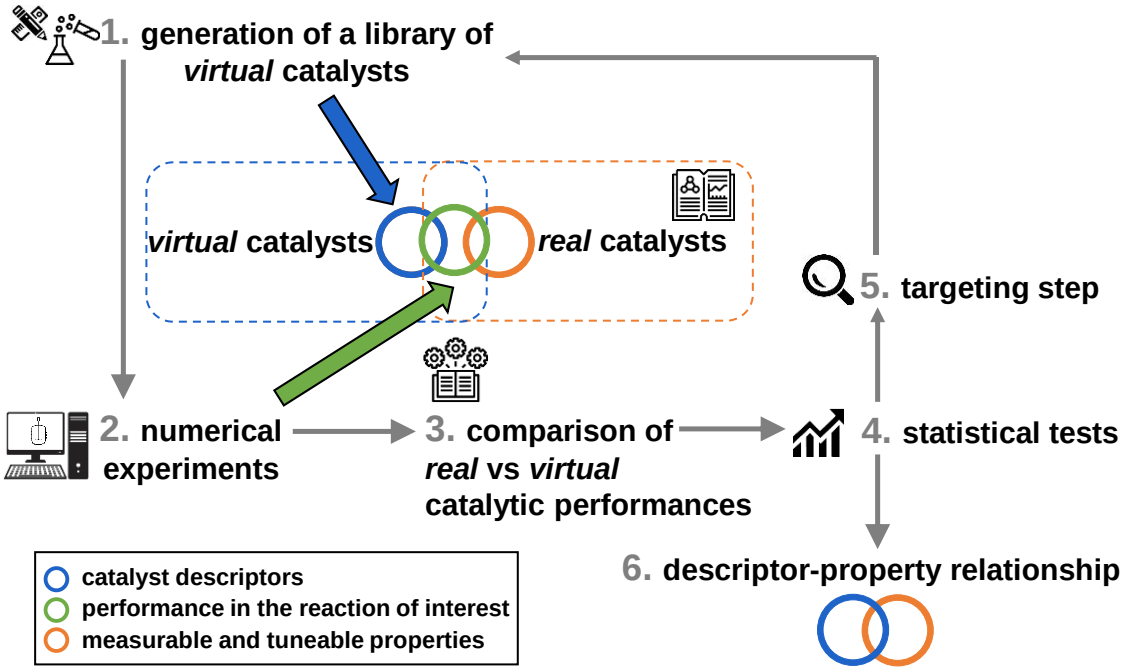


Figure 1. Schematic overview of the kinetics-driven design methodology . The central element of the methodology is the comparison of experimental and simulated data. Virtual catalysts are described by a set of catalyst descriptors (blue) and the simulated performances (green) associated to these descriptors. Real catalysts are described via a measurable property (orange) and experimental performances (green) reported in literature. The thin, grey arrows indicate the order of the methodological steps; the thick, coloured arrows indicate the source of the data required for this comparison. The icons used in this figure are described in Table 1.

The "philosophy" behind the relationship between the properties of a catalyst and its descriptors is based on matching, i.e. identifying similarities in experimental (*real*) and simulated (*virtual*) performances (*Step 3* in Figure 1), and on assessing how the descriptors and the properties of the matching *virtual* and *real* catalysts are related (*Steps 4 and 6* in Figure 1). In practice, several consecutive steps are followed. These are schematically represented in Figure 1 and are described as follows:

Step 1. In the first iteration of the cycle, a *discovery* library of *virtual* is generated, while in subsequent iterations *targeted* libraries (see *Step 5*) are constructed.

Step 2. The *virtual* catalysts from *Step 1* are used as input to a fundamental kinetic model integrated in a reactor model. The performances of the *virtual* catalysts are simulated at process conditions corresponding to the dataset of *real* catalysts under investigation.

Step 3. The experimental (from the selected dataset) and simulated performances (from *Step 2*) are compared via k-means clustering[54], in order to establish which *virtual* catalysts approach the performance of the *real* ones;

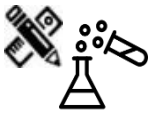
Step 4. The important descriptors are those that discriminate different clusters, and are identified via statistical tests on catalyst descriptors distributions;

Step 5. If the *discovery* library from *Step 1* does not result in a statistically significant number of matches, a *targeted* library which is refined over the consecutive iterations if needed. In this targeting step, the ranges for the discriminating descriptors identified in *Step 4* are narrowed down. The methodology is then iterated starting from *Step 1*.

Step 6. The iterative cycle is considered completed when the *targeted* library is sufficient to establish a statistically relevant relationship between descriptors and measurable and tuneable properties.

The icons presented in Figure 1 are explained in Table 1 together with a list of the tools used in this study. Each of the above mentioned steps and tools will be concisely described in the next sections. The reader is referred to the cited literature for more details on the applied techniques.

Table 1. Tools used in the present work. The icons in the first column are the same as in Figure 1.

	Meaning	Tool
	literature data	manual web search
	space-filling design of (numerical) experiments; in the present work: FFF (Fast Flexible Filling)	JMP® 13.2.1[55]
	numerical experiments: kinetics-based simulations	in-house code[56]
	k-means clustering in the space of catalytic performances	Orange3[57]
	Kruskal-Wallis/ Mann-Whitney tests on descriptors distributions	JMP® 13.2.1[55]
	narrowing descriptor ranges; in the present work: according to Chebyshev's inequality	JMP® 13.2.1[55] + non-automated iteration

Step 1. Generation of a library of *virtual* catalysts

n virtual catalysts are represented each by a unique set of m descriptors resulting in an $n \times m$ matrix. The catalyst descriptors and their value ranges thus determine the *virtual* catalysts design space. This space is sampled via an adequate Design of Experiments (DoE) technique to generate a diverse library of *n virtual* catalysts, each represented by a combination of the m catalyst descriptors. These key elements have been schematically represented in Figure 2 and are discussed in some more detail below.

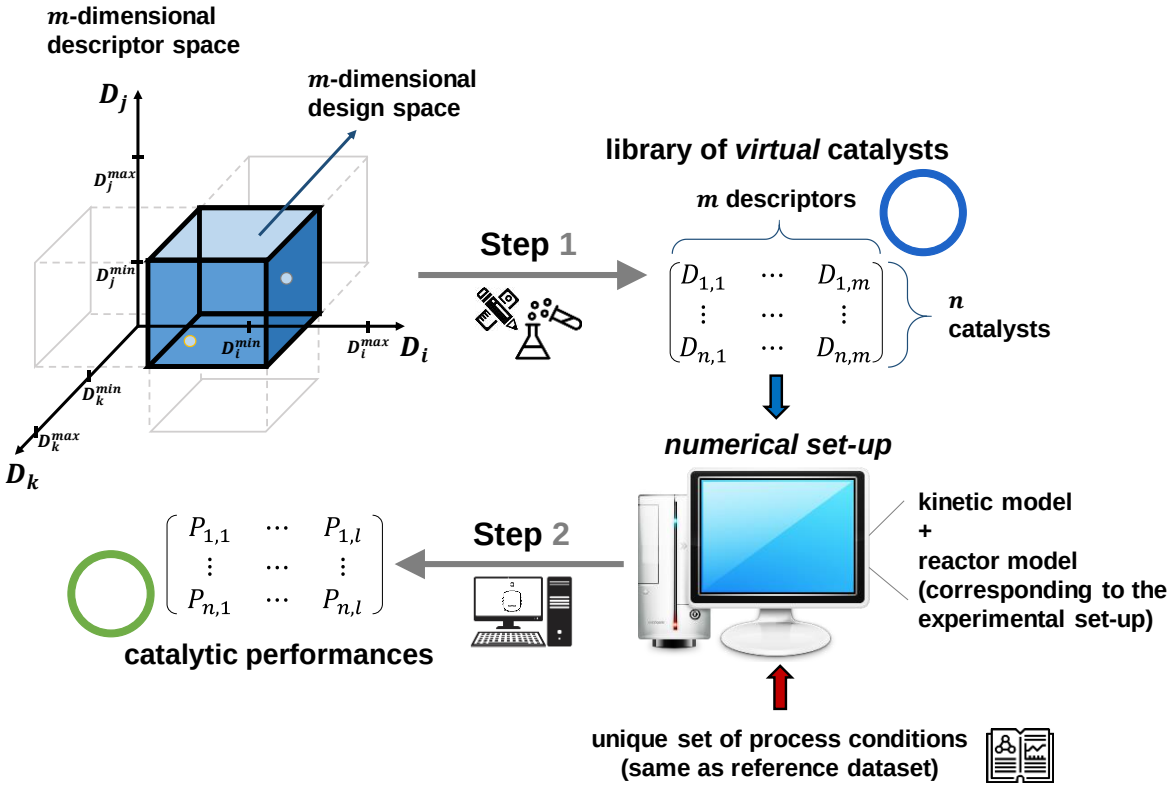


Figure 2. Visual representation of methodological steps 1, namely the generation of a library of virtual catalysts, and 2, i.e. the numerical testing of the virtual library. In the figure, the library is shown in its matrix form. The same colour scheme is applied throughout the whole section: blue refers to catalysts descriptors, green to catalytic performances.

Catalyst descriptors D_j are parameters in the fundamental kinetic model, and their number and type are highly specific to the complexity of the studied reaction. In fact, the number of descriptors can increase rapidly with the number of species and reaction steps considered, and the types of active sites included. An example of a set of catalyst descriptors for OCM is provided in section 3.

The value range of each descriptor defines the boundaries of the m-dimensional design space in which the n virtual catalysts are generated. The range of descriptors is the broadest for the initial *discovery* library, and is preferably based on physico-chemical considerations and values reported in relevant literature. For the *targeted* libraries, the ranges of the discriminating descriptors are narrower than the original ones, see *Step 5*.

To limit the computational time in *Step 2*, n has to be adequately chosen. DoE techniques for computer simulations, or. 'numerical' experiments[58-60], herein denoted as Do(N)E, are applied

for this purpose. The most suitable technique depends on the number and the nature (qualitative – quantitative, continuous – discrete) of descriptors, which are denoted as *factors* in Do(N)E terminology, as well as the number of values or '*levels'* associated with each *factor*[58]. Space-filling designs[61] are herein of interest thanks to their good applicability to deterministic systems with a high number of *factors*. In the present work, the Fast Flexible Filling (FFF) method[62] was selected and implemented via the statistical software JMP® 13.2.1[55]. This methodology is recommended because, in contrast to other space-filling designs, the FFF design space does not have to be a hypercube: constraints can be applied in the design space and, hence, non-rectangular projections of the design space are possible. Furthermore, FFF can work also with discrete *factors*, such as, for catalysis, the framework type and nature of the active site (e.g. Brønsted acid (0) or metal (1), discrete values in brackets);

The number of design points (n) for each *factor* is defined from a typical rule of thumb for space-filling designs: $n = 10 \times m$[63]. If deemed necessary, an improved design can be created by increasing n.

Step 2. Numerical experiments

The resulting *n virtual* catalysts are simulated for their performance in the reaction of interest using a so-called *numerical set-up.* The *numerical set-up* consists of a validated kinetic model embedded in a reactor model. It is highlight here that this model represents a pre-requisite for the present methodology.

Figure 2 represents a *numerical set-up* as a generic computer. For each *virtual* catalyst i tested in the *numerical set-up*, l performance indicators $P_{i,k}$ can be obtained, such as conversions, selectivities, turnover frequencies, catalyst lifetimes, etc.

The performance of the n *virtual* catalysts are compared with those of *real* catalysts by adding the experimental data to the simulated ones. It is important that the conditions at which the performances are simulated correspond with the ones reported for the experimental dataset. If

the reference *real* catalysts have been tested at several process conditions, the whole methodology should be repeated for all sets of process conditions, and the presence of a kinetic model at the heart of the *numerical set-up* allows for that.

The dataset acquired on *real* catalysts should comprise catalysts that differ both in performance and in at least one property which can be easily measured via physico-chemical characterization. The size of the dataset is then of lesser importance; even more, small (< 10 samples) datasets, which are typically of no use in advanced machine learning applications, can be more easily investigated with the proposed methodology.

Step 3. Comparison of *real* vs *virtual* catalytic performances

The n *virtual* performances and q *real* performances are compared and matched using a clustering algorithm in the l-dimensional space of the performance indicators. k-means[54] clustering is an unsupervised machine learning technique that discovers hidden patterns or structures in data[64]. Data are grouped in clusters based on similarity which corresponds to a distance (in this case Euclidean distance): the lower the distance between two objects, the higher the similarity and vice versa[15].

The free Python application Orange3[57] was selected for the analysis. This software effectively deals with the main issues of k-means clustering: the difficulty of selecting a priori a suitable number of clusters, via silhouette scoring[65], and the high sensitivity of the result to the initialization phase of the algorithm, via the k-means++ initialization algorithm[66].

The clustering algorithm assigns each catalyst, *virtual* or *real*, to one of the N_c clusters according to the performance indicators. This step of the methodology is graphically summarized in Figure 3.

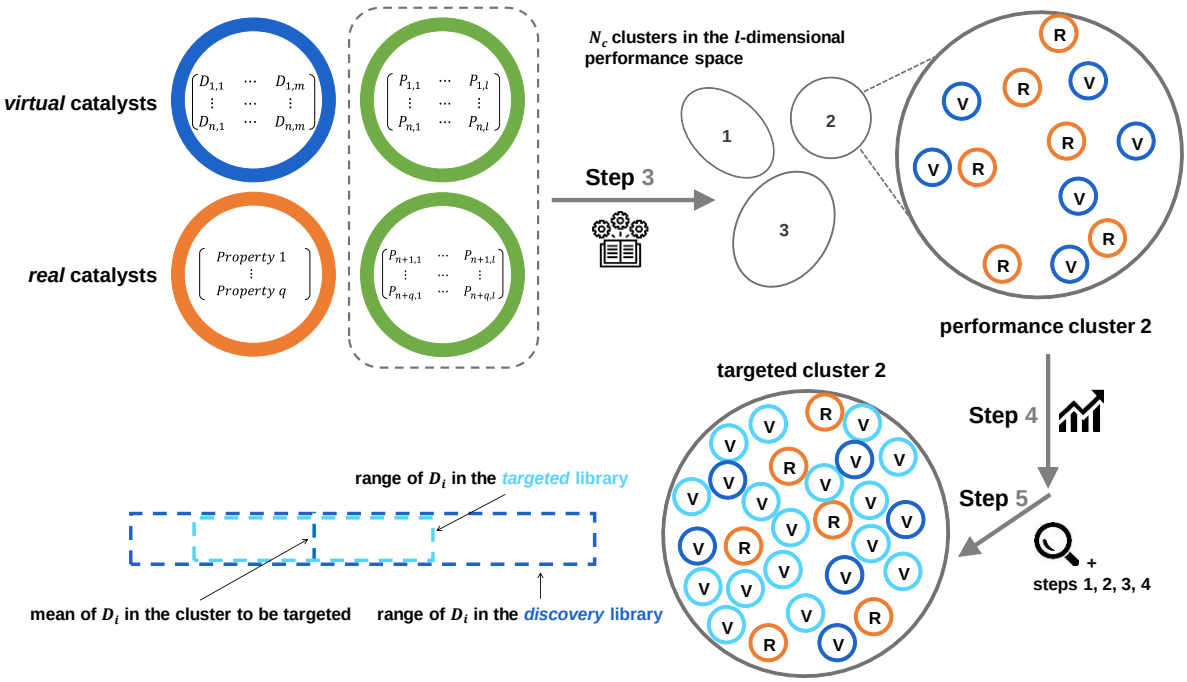


Figure 3. Visual representation of methodological steps 3 and 5. A simplified example of a clustering procedure (Step 3), applied to both virtual and real catalysts in the space of catalytic performances, and of the targeting procedure (Step 5), applied to a generic performance cluster 2 via descriptor D_i, is provided. The intermediate Step 4 is depicted in Figure 4. R indicates real catalysts, V indicates virtual catalysts. The same colour scheme is applied throughout the whole section: blue refers to catalyst descriptors, green to catalytic performances, orange to catalysts properties.

Step 4. Statistical analysis of descriptor distributions

For each cluster, m probability distributions describe the variability of each descriptor in this cluster. The comparison of these distributions among different clusters indicates which descriptors allow discriminating between clusters and thereby the performances of *virtual* catalysts. The concept of discrimination is exemplified in Figure 4 by distributions of descriptors D_i and D_j. Distributions D_i^1 and D_i^3 are statistically different from one another to pinpoint D_i as a 'discriminating' descriptor, unlike descriptor D_j. As will be described later on in section 6, in some cases this discriminating descriptor can be related to a catalyst property, as suggested in the figure.

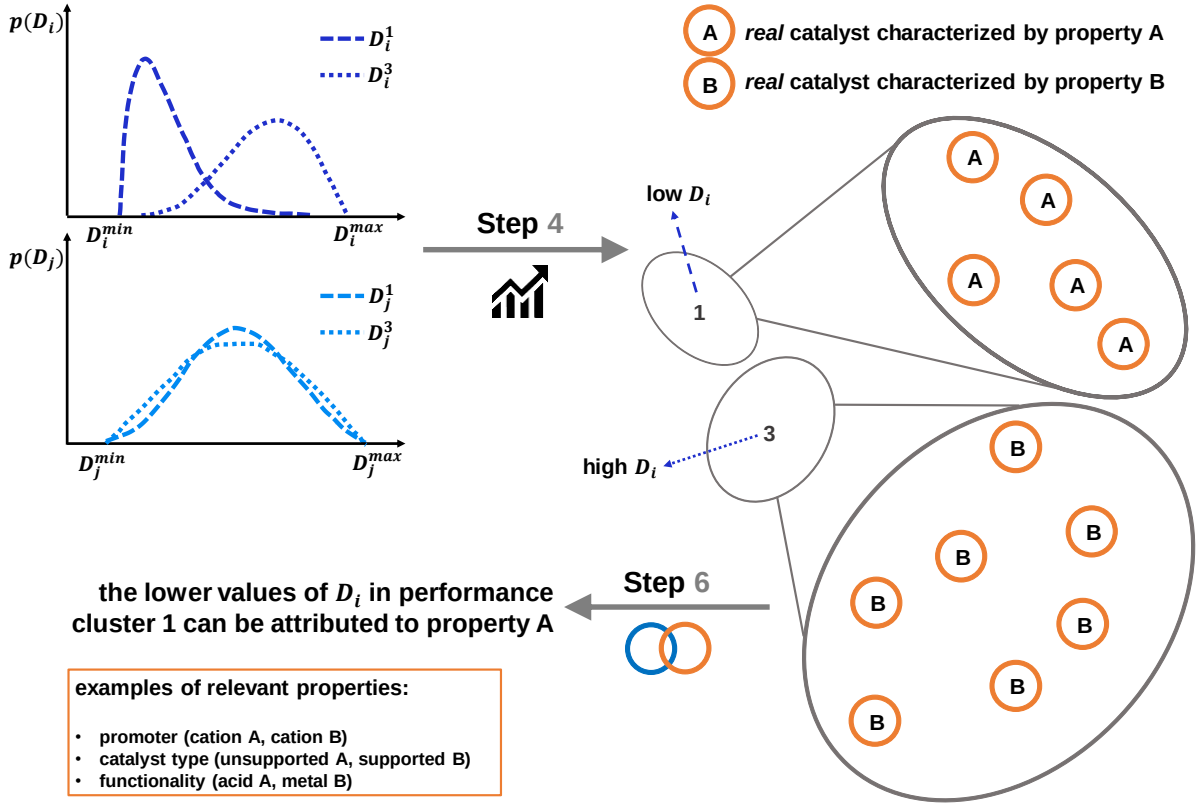


Figure 4. Visual representation of methodological steps 4 and 6. A simplified example of the comparison of descriptors' distributions between different clusters (Step 4) and of descriptor-property relationships which can be established with the proposed methodology (Step 6) is provided. The same colour scheme is applied throughout the whole section: blue refers to catalyst descriptors, orange to catalysts properties.

The descriptors distributions are considered 'sufficiently' different using specific statistical tests. In conventional or *parametric* statistics the data are assumed to fit a pre-defined distribution[67], commonly the normal distribution. The difficulty, however, is that catalyst descriptors within a cluster are typically not normally distributed. Thus, *non-parametric* tests[68] are preferred. The statistical tests proposed here are the Kruskal-Wallis[69] and the Mann-Whitney[70] tests, performed via the software JMP® 13.2.1[55]. These tests are *non-parametric* alternatives to, respectively, the ANOVA test and the t-test of Student. The former can be used for a number of clusters $N_c > 2$ and indicates whether there is at least one cluster with a significantly different distribution of each descriptor. The Mann-Whitney test compares the descriptor distributions between two clusters ($N_c = 2$). These tests have less statistical power

than *parametric* tests and great care should be taken in the interpretation of results. For that reason, the Mood's median test[71] was applied as a secondary verification tool. The theoretical background of the mentioned tests can be found in the above references and in the manual of the statistical software[55].

Step 5. *Targeted* libraries of *virtual* catalysts

The number of *virtual* catalysts per cluster may vary significantly. To establish statistically significant relationships between descriptors and relevant properties in *Step 6*, the number of *virtual* catalysts in the proximity of the *real* catalysts is increased. This is achieved by generating a *targeted* library of *virtual* catalysts per cluster.

The targeting procedure, earlier represented in Figure 3, consists of defining progressively narrower value ranges while typically maintaining the mean value. The procedure is only applied to the discriminating descriptors from *Step 4*. With those new ranges, library generation (*Step 1*), numerical testing (*Step 2*) and clustering are iterated (*Step 3*), until a 'sufficient' number of *virtual* catalysts closely matches the *real* catalysts in performance. There is no unambiguous definition of 'sufficient' as it greatly depends on the type of analysis one is interested in. A general criterion is to perform iterations until the number of *virtual* catalysts in the target cluster is no longer increasing.

During the first iteration, new descriptor ranges were herein chosen as a band of $\pm\sqrt{2}$ standard deviations around the mean of the original cluster from the *discovery* library: $D_i|_{targeted}^{C_a} = \overline{D_i}|_{discovery}^{C_a} \pm \sqrt{2}\sigma_i|_{discovery}^{C_a}$. According to Chebyshev's inequality[72], this formula narrows down the descriptor range while retaining at least 50% of *virtual* catalysts within the target cluster. During successive iterations, the means and the standard deviations are referred to the *targeted* library generated in the previous iteration.

While not applied in this work, the targeting procedure can be performed automatically using any kind of evolutionary algorithm[73] or self-organizing map[74].

Step 6. From properties of *real* catalysts to descriptor-property relationships

This step highly depends on the investigated catalytic reaction and experimental dataset. A few examples are given for OCM later on, while Figure 4 presents a more generic case. Clusters 1 and 3 in the figure contain *real* catalysts which exhibit reported differences in a specific property (e.g. composition, type of support, etc.), denoted as A or B depending on the cluster. These differences can now be related to the discriminating descriptors from *Step 4*, in this example, descriptor D_i. Therefore it can now be suggested that a certain property having value A, or not having value B, affects the values of D_i and, via it, the performances. At this point, a qualitative relationship has been established between a measurable and tuneable property of a catalyst and the descriptor which portrays the catalyst role within the kinetics. Further analysis of the dataset and the libraries can lead to other fundamental insights in terms of descriptor-property relationships, as will be described in the next section.

3. Application to the Oxidative Coupling of Methane

The methodology from Section 2 was herein applied to three OCM datasets from literature. These datasets were selected for the completeness of the information reported and because, in all three cases, the authors experimentally verified the impact of a relevant catalyst property on the OCM performance.

Steps 1 to 3 were similar for all three datasets, as the same fundamental kinetic model for OCM[56], previously developed in our research group, was applied. This model accounts for 39 reversible gas-phase reactions and 26 reversible surface reactions, which are reported in Tables S1 and S2 of the Supporting Information, and introduces the m= 16 catalyst descriptors listed in Table S3 of the Supporting Information. A detailed explanation of how these descriptors impact the kinetic parameters is provided in previous literature[50] and a summary is provided in Table S4 of the Supporting Information.

- *Step 1. Generation of a discovery library of virtual catalysts* A *discovery* library of $n = 20 \times m = 320$ *virtual* catalysts was generated by applying the FFF design to the 16 catalyst descriptors, within the feasibility ranges retrieved from literature and reported in Table S3 of the Supporting Information. A 2D visualization of the resulting 16-dimensional library is reported in the Supporting Information (Figure S1). The same *discovery* library was used for the three datasets.
- *Step 2. Numerical experiments* The *numerical set-up* used for simulating performances of the *virtual* catalysts consists of the earlier mentioned kinetic model embedded in a 1-dimensional, heterogeneous reactor model, i.e. not including radial gradients on the reactor scale, but including transport limitations on the particle scale[75]. For each of the three case studies, a single set of process conditions was considered, that corresponds to the selected experimental dataset.
- *Step 3. Comparison of real vs virtual catalytic performances* Methane conversion, X_{CH4}, and selectivity to C_2 hydrocarbons (C_2H_6, C_2H_4, C_2H_2), S_{C2}, were selected as two (l= 2) performance indicators for all case studies, as both are crucial in reaching an economical feasible product yield.

Other aspects of *Steps 2 and 3*, together with *Steps 4-6*, are specific to each dataset. These are described in detail for the first dataset as a comprehensive example. For the second and third dataset the focus is mainly on *Step 6*, to highlight the descriptor-property relationships that could be established.

3.1 Dataset 1: diversity in composition

For this first case study, the dataset from Kondratenko et al.[47] was selected. The property of interest in this dataset is the composition, which is reported for a large number of *real* catalysts. The authors synthesized and evaluated q= 44 three-component catalysts with La_2O_3 or MgO as host metal oxide, and two promoters which are oxides of Li, Na, Cs, Sr, Ba, La or Mn

(concentrations from 0 to 10 wt%). A list of these catalysts can be found in Table S5 in the Supporting Information and the process conditions are reported in Table S6. The corresponding experimental performances are shown in Figure 5/A in terms of S_{C2} as function of X_{CH4}. In the same graph, iso-C_2 yield (yield= conversion × selectivity) curves are also shown. It can be observed that none of the *real* catalysts (indicated by full triangles) in this dataset exceeds 20% yield.

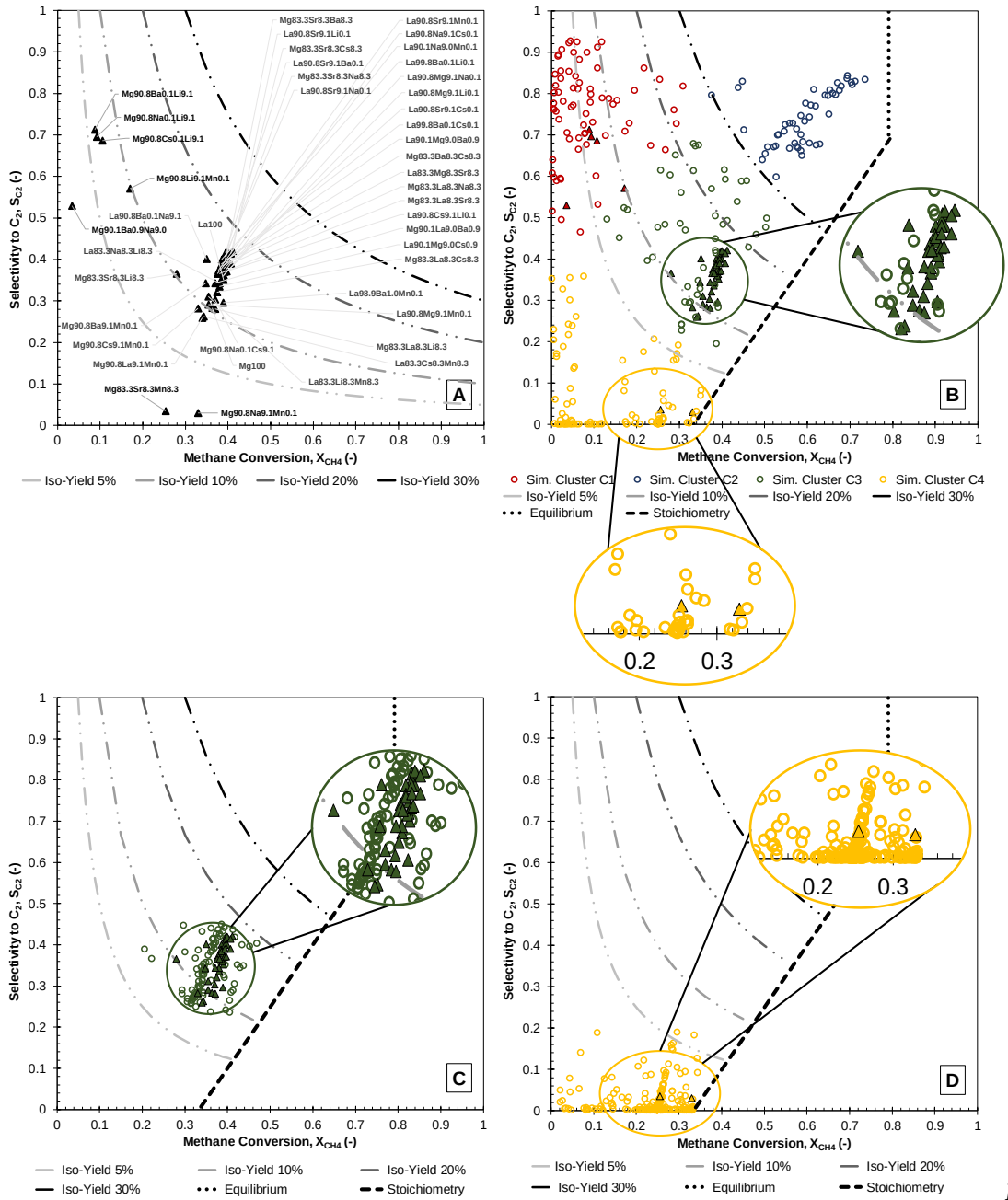


Figure 5. A. Real catalysts (full triangles)= experimental data from the dataset of Kondratenko et al.[47]; B. Real (full triangles) + virtual catalysts (empty circles)= output of the numerical testing of the virtual catalyst library (Step 2); the colours distinguish the four clusters of real and virtual catalysts, obtained with the k-means algorithm

implemented in the Orange software[57] (Step 3). Regions with real catalysts are zoomed in; C. Output of the targeting procedure (Step 5) for cluster 3; D. Output of the targeting procedure (Step 5) for cluster 4. Unique set of process conditions: T= 1073 K, p= 1 bar, $W/F_{CH4,0}$= 71 kg_{cat} s/$mol_{CH4,0}$, CH_4/O_2= 2.

The authors observed that the promoters did not affect significantly the performances of La_2O_3-based catalysts. The highest C_2 yields of approx. 17% were achieved over 0.1%wtLi9.1%wtSr/La_2O_3 and 8.3%wtCs8.3%wtSr/MgO. The beneficial effect of Sr was confirmed by the authors via regression tree analysis and it was tentatively attributed to the formation of anion vacancies in the lattice of the host oxide. They also identified the presence of Mn as detrimental for the performances of both MgO and La_2O_3 catalysts, but no clear explanation was provided.

<u>*Step 2. Numerical experiments*</u> The performances of the *virtual* catalysts are displayed in Figure 5 in addition to the experimental performances of *real* catalysts. In Figure 5/B two lines are included: the vertical line represents the 'equilibrium' line, while the diagonal line corresponds to the 'stoichiometry' line. These lines define the boundaries for realistic performances in the conversion-selectivity plane. The equilibrium line indicates a maximum equilibrium methane conversion of 0.79 at the set of process conditions, assuming S_{C2H4}= 1. The stoichiometry line corresponds to the maximum methane conversion that can be obtained at complete oxygen conversion. For CH_4/O_2= 2, the maximum methane conversion increases from X_{CH4}= 0.33 for only partial oxidation to CO ($CH_4/O_2|_{stoich}$= 2/3), to X_{CH4}= 1 for coupling to ethylene, i.e. S_{C2H4}= 1 ($CH_4/O_2|_{stoich}$= 2).

The *discovery* library preferably covers a broad range of catalytic performances, comprising diverse activities and selectivities within chemically feasible boundaries (Figure 5/B). In addition, the performances of all *real* catalysts were approached by several performances of *virtual* catalysts, enabling the matching procedure from step 3.

<u>*Step 3. Comparison of real vs virtual catalytic performances*</u> The results from the *clustering* algorithm is shown in Figure 5/B in different colours. The Orange3 software identified four

clusters (N_c= 4) as the optimal subdivision of the performance space. This value was also in line with visual intuition. In fact it can be observed that the four clusters, numbered clockwise starting from the top left, are indicative of different C_2 yield levels: medium-low (C1), high (C2), medium-high (C3) and low (C4). The data distribution over the clusters is described in Table 2.

Table 2. Description of the clusters obtained in the performance-based comparison via k-means clustering (Step 3) applied to the dataset of Kondratenko et al.[47] and reported in Figure 5/B..

Cluster	**Colour**	**Number *virtual* catalysts**	**Number *real* catalysts**	**X_{CH4} [%] (mean and standard deviation)**	**S_{C2} [%] (mean and standard deviation)**	**C_2 Yield [%] (mean and standard deviation)**
C1	Red	84	5	8.2 ± 7.6	75.0 ± 11.2	6.2 ± 5.8
C2	Blue	60	0	59.1 ± 7.6	74.0 ± 6.6	44.0 ± 8.6
C3	Green	60	37	35.5 ± 6.8	41.0 ± 11.4	14.5 ± 5.0
C4	Yellow	116	2	14.4 ± 11.1	6.4 ± 9.6	0.7 ± 1.2

To relate *real* and *virtual* catalysts, the only clusters of interest were those containing both. Hence, C2, despite the high yield simulated, was the least significant one for further analysis.

Step 4. Statistical analysis of descriptor distributions Figure 6 depicts the distribution of two descriptors, namely the logarithms of D10, sticking coefficient of O_2 (left), and D11, sticking coefficient of methyl radicals (right), in the four clusters identified in *Step 3*, in terms of probability density functions (A) and box plots (B). These two graphical representations were chosen to demonstrate that the descriptors are non-normally distributed within each cluster. Such behaviour is believed to be due to the performance being a complex, highly non-linear function of both the process conditions and the sixteen catalyst descriptors. The analysis of the D10 distributions suggested no significant differences among the four clusters. Descriptor 11, on the other hand, seemed to exhibit some cluster-specific trends, with lower values in C1 (high yield) and higher values in C4 (low yield).

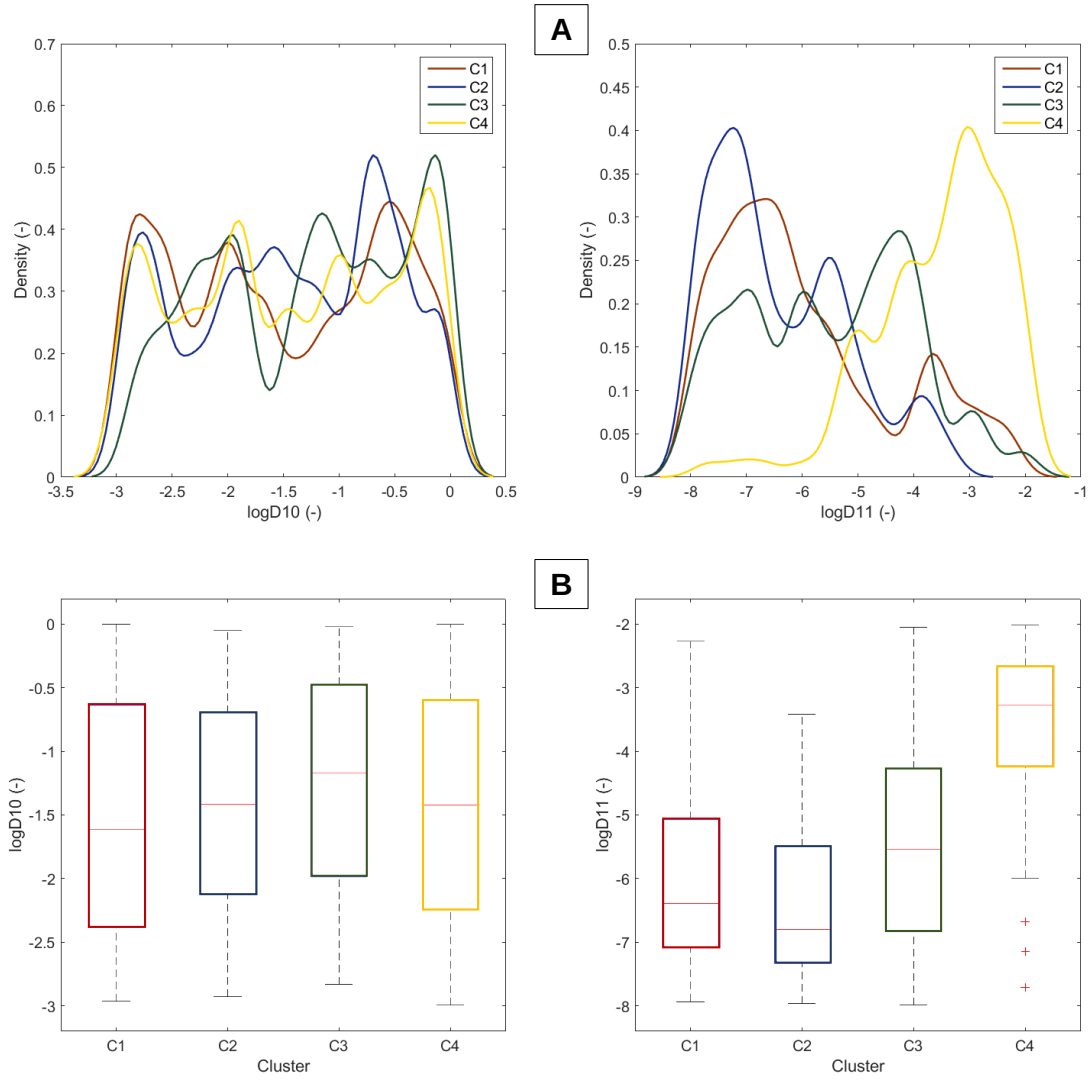


Figure 6. Comparison of the distributions of descriptors logD10 and logD11 in the four clusters shown in Figure 5/C, in terms of probability density functions (A) and box plots (B). On each boxplot, the central red line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The dashed whiskers extend to the datapoint closest to +/– 2.7 standard deviations, and the datapoints which do not belong to this interval are plotted individually using the '+' symbol. As an example, the three '+' points in the right graph of figure B, for cluster 4, correspond to the small shoulder in the corresponding distribution (in yellow) in the right graph of figure A, between logD11= -8 and logD11= -6.

After this preliminary analysis, the observed trends were statistically verified for all sixteen descriptors. A complete analysis is reported in the Tables S7 and S8 of the Supporting Information. Both Kruskal-Wallis and Mood's Median tests gave comparable results and the discriminating descriptors are, in order of statistical relevance: D11 (sticking coefficient of CH_3 on a O* species on the catalyst surface), D1 (reaction enthalpy of H-abstraction from CH_4 by O*), D16 (density of active sites), D15 (sticking coefficient of C_2H_4 on O*). D10 was discarded

as was previously observed in Figure 6. The impact of the discriminating descriptors on the kinetics is more extensively described in section S1 of the Supporting Information.

Step 5. Targeted libraries Figure 7 reports the distribution of the four discriminating descriptors as previously identified, for the original *discovery* library and two *targeted* libraries for C3 and C4. The descriptor value ranges were reduced during the targeting with respect to the *discovery* library, and statistically significant differences (with 95% probability) appeared when comparing the two *targeted* libraries.

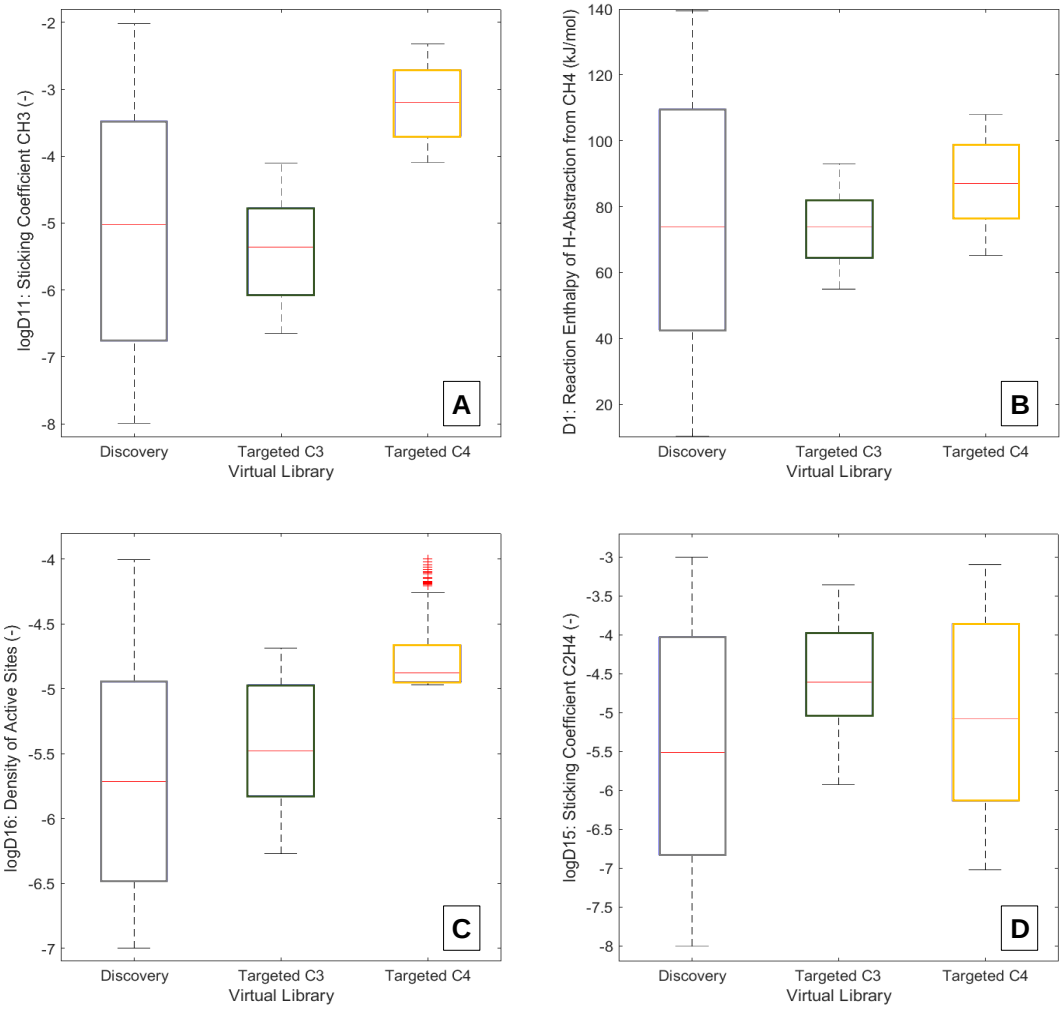


Figure 7. Comparison of the distributions of log D11: sticking coefficient of CH_3· (A), D1: reaction enthalpy of H-abstraction from CH_4 (B), logD16: density of active sites (C) and logD15: sticking coefficient of C_2H_4 (D) in the libraries of virtual catalysts which were generated and tested in order to target the performances of clusters 3 and 4 of the dataset of Kondratenko et al.[47].

Figure 5/C and Figure 5/D show the library of data points as a result of the targeting procedure applied to C3 and C4, by using the new descriptor ranges from Figure 7. As desired, the

narrowing of the descriptor ranges resulted in narrower performance ranges. More importantly, without having increased the total number of catalysts in the library, the distance in performance between *virtual* and *real* catalysts was closed down. The targeting was therefore considered successful.

Step 6. From properties of real catalysts to descriptor-property relationships From Figure 7 it was observed that the *targeted* library for C4 (low yield) exhibited higher values of D11, compared to the library targeting C3 (medium-high yield). The high D11 implies a larger contribution of catalytic oxidation of methane (methyl radicals stick and are oxidized on the catalyst surface), causing the low C_2 selectivity in C4. D1, which is inversely related to methane activation, is slightly higher in C4, while the active site density D16 is also higher (more oxidation sites available). Despite being statistically significant, no visible differences in the distributions of D15 could be observed.

To establish the link with the catalyst compositions, one could observe that the *real* catalysts in C4 both contain MgO as host oxide and Mn as promotor (see Figure 5/A and D). This somewhat agreed with the worst performing samples in C3 containing Mn, as identified by the authors originally. However, only two *real* samples are present in C4 and, more importantly, some samples which include both MgO and Mn are also present in C3. Hence, the hypothesis that the presence of Mn in a MgO matrix impacts the four discriminating descriptors and, consequently, the C_2 selectivity could not be statistically verified. On the other hand, the *real* catalysts in C3 were characterized by a very high diversity in composition and very limited diversity in performance. This means that, even if some cluster-specific composition characteristics could be qualitatively observed, the lack of statistical validation precluded the establishment of these suggestive descriptor-composition relationships.

3.2 Dataset 2: diversity in surface basicity

In this second case study, the methodology was applied to the experimental dataset of Kuś et al.[76]. The property of interest is the strength of basic sites. The authors of this work examined the effect of surface basicity on the catalyst performance by synthesizing and testing four unmodified pure metal oxides: La_2O_3, Nd_2O_3, ZrO_2 and Nb_2O_5. For each of the first three oxides, two samples were prepared by varying the calcination atmosphere (O_2 for sample a, and N_2 for sample b). The authors characterized the *real* catalysts via routine techniques and identified La_2O_3 as the catalyst with the strongest basic sites followed by Nd_2O_3, and both catalysts exhibited a good OCM performance. ZrO_2 is amphoteric, but with prevailing basicity. Nb_2O_5 is also amphoteric but is rather a solid acid catalyst. The calcination atmosphere was found to have a negligible influence on the basicity and is not further considered. An overview of the data and the corresponding set of process conditions is given in section S4 of the Supporting Information. The reader interested in impact of catalyst basicity on C_2 yield is referred to dedicated OCM literature, such as e.g. the work of Rane et al.[77].

Steps 1 and 2 are analogous as from the previous case study; the process conditions used in the numerical experiments were adapted to those reported by Kuś et al.[76].

Step 3. Figure 8 shows the results obtained from the performance-based clustering procedure. The procedure is detailed in Table S11 of the Supporting Information. Four clusters were identified and numbered clockwise starting from the top-left. It can be observed that *real* catalysts are situated in cluster 3 (medium yield, green) and cluster 4 (low yield, yellow). The orange arrow in the figure, indicates the direction of growing strength of basic sites, as previously identified by the authors of the reference work. This suggests that the property shared by the *virtual* catalysts belonging to C3 is a higher strength of basic sites, as opposed to the catalysts belonging to C4.

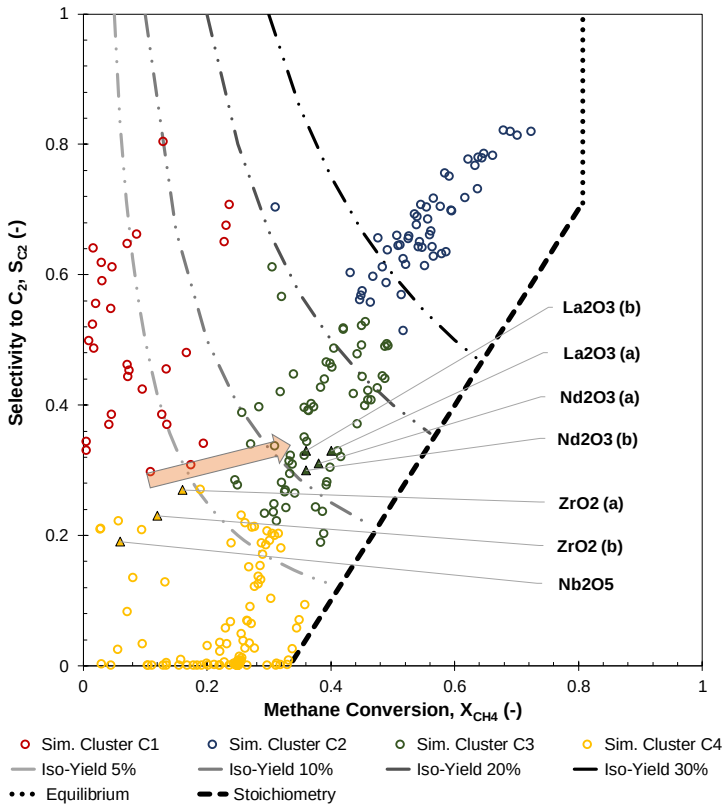


Figure 8. Four clusters of real (full triangles) and virtual (empty circles) catalysts, obtained with the k-means algorithm implemented in the Orange software[57] for the dataset of Kuś et al.[76]. The colours distinguish the four clusters. Unique set of process conditions: T= 1033 K, p= 1 bar, $W/F_{CH4,0}$= 1031 kg_{cat} s/$mol_{CH4,0}$, CH_4/O_2= 2. The arrow indicates the direction of increasing basic strength of the active sites.

Step 4. The Mann-Whitney test followed by a verification with Mood's median test were applied to the two clusters of interest to identify the discriminating descriptors; see section S4 of the Supporting Information for all details. D1 (reaction enthalpy of H-abstraction from CH_4) and D2 (chemisorption enthalpy of O_2), were found to be the discriminating descriptors.

Step 5. The ranges of these descriptors were re-considered to generate *targeted* libraries for La_2O_3 (a and b) and Nd_2O_3 (C3), and for ZrO_2 (a and b) and Nb_2O_5 catalysts (C4). The distributions of these descriptors in the two *targeted* libraries are shown as box plots in Figure 9.

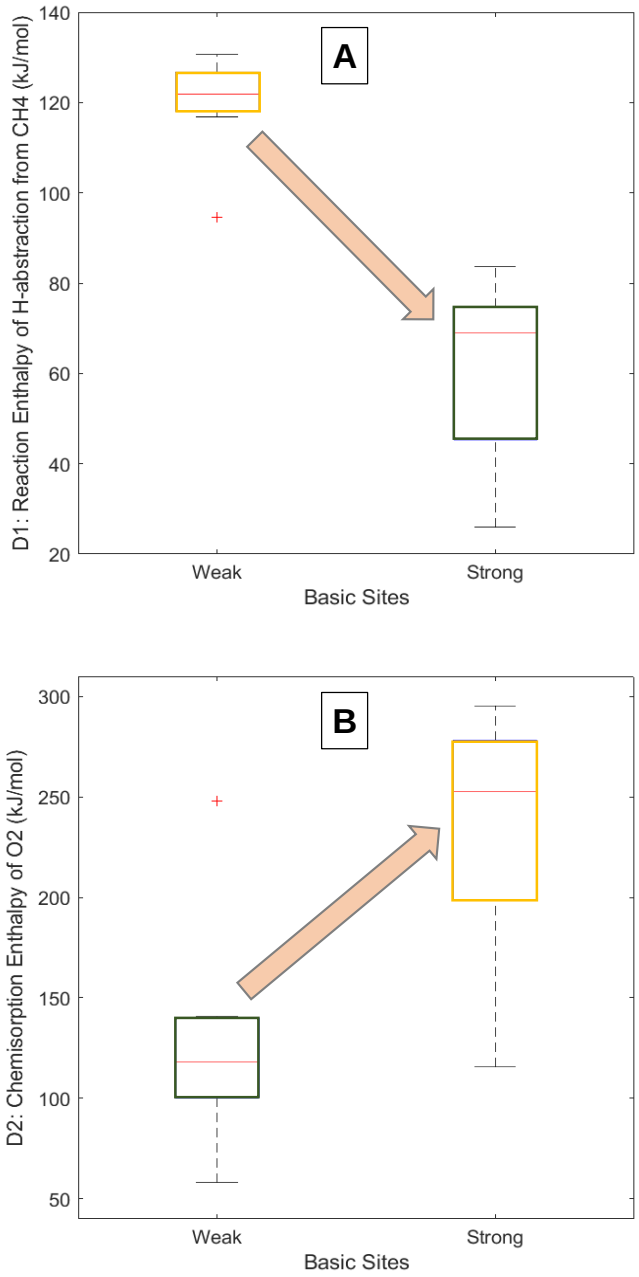


Figure 9. Comparison of the distributions of D1: reaction enthalpy of H-abstraction from CH_4 (A) and D2: chemisorption enthalpy of O_2 (B) in the targeted libraries around C3 (strong basic sites) and C4 (weak basic sites) from the dataset of Kuś et al.[76] *with weak (C4) or strong (C3) basic sites. The arrow indicates the direction of growing strength of basic sites.*

Step 6. From properties of real catalysts to descriptor-property relationships In Figure 9, the boxplot distributions do not overlap between both *targeted* libraries, with the *virtual* catalysts from C3 having a lower D1 (i.e. easier methane activation). The *targeted* library for C3 is also characterized by a higher oxygen chemisorption enthalpy (medium-high D2). Both catalyst

descriptors impact the catalyst performance and could be clearly related to the basicity, hence, establishing two descriptor-basicity relationships.

An easier methane activation (lower D1) for catalysts with a high surface basicity (C3) has already been proposed in the past[49, 78, 79] and was for the first time confirmed via a detailed reaction mechanism and validated by statistical tests herein. D2 and the strength of surface basic sites are related to the nature of surface oxygen species. Stronger oxygen chemisorption occurs in the form of lattice oxygen (O^{2-}), while materials that bind oxygen more weakly (lower D2) rather contain other charged oxygen species (O^{-},O_2^{-},O_2^{2-})[39]. Lattice oxygen is more nucleophilic and characterized by higher basicity[80]. On the other hand, counterintuitively, the density of basic sites (D16) does not seem to be majorly impacted by the differences in basicity. This might be related to the fact that both an increase in D16 and a decrease in D1 contribute to faster methane activation. The effect of D16 might, hence, be partly covered by the strong effect of D1 during the statistical tests.

3.3 Dataset 3: diversity in electronic properties

In the dataset of Malekzadeh et al.[81], the electronic properties of OCM catalysts were studied. More precisely, the authors sought for a correlation between electronic properties and catalytic performance for 4%wt metal oxides supported on Na_2WO_4/SiO_2 catalysts. Seven catalysts were prepared and tested, both unpromoted Na_2WO_4/SiO_2 and promoted by various metals: V, Cr, Mn, Fe, Co or Zn. The authors characterized the *real* catalysts in terms of conductivity and band gap[81]: the best performing catalyst, $Mn/Na_2WO_4/SiO_2$, is a p-semiconductor and exhibited high electrical conductivity, and low band gap; $Co/Na_2WO_4/SiO_2$, $Fe/Na_2WO_4/SiO_2$, $Zn/Na_2WO_4/SiO_2$, and Na_2WO_4/SiO_2, which gave intermediate C_2 yields, also exhibited intermediate electrical conductivity and band gap and can be classified as both p-type and n-type semiconductors; the least performing catalysts, $Cr/Na_2WO_4/SiO_2$ and $V/Na_2WO_4/SiO_2$, presented lower electrical conductivity and a higher band gap. An overview of the data and the

corresponding set of process conditions is given in section S5 of the Supporting Information. The reader is referred to dedicated OCM literature for more information on the impact of electronic properties, such as the work of Zhang et al.[82].

Steps 1 and 2 described in the first case study; the process conditions used in the numerical experiments were adapted to the ones considered in the experiments performed by Malekzadeh et al.[81].

Step 3. Figure 10 shows the results obtained after the performance-based clustering procedure. The corresponding description is reported in Table S16 of the Supporting Information. Again, four clusters were identified and numbered clockwise starting from the top-left. It can be observed that *real* catalysts are present in C2 (high yield, blue), C3 (medium-high yield, green) and C4 (low yield, yellow). The orange arrow in the figure indicates the direction of growing electrical conductivity. *Step 4.* Kruskal-Wallis test and Mann-Whitney tests, followed by a verification with Mood's median test, suggested two discriminating descriptors: D11 (sticking coefficient of $CH_3\cdot$), and D1 (enthalpy of H-atom abstraction from CH_4).

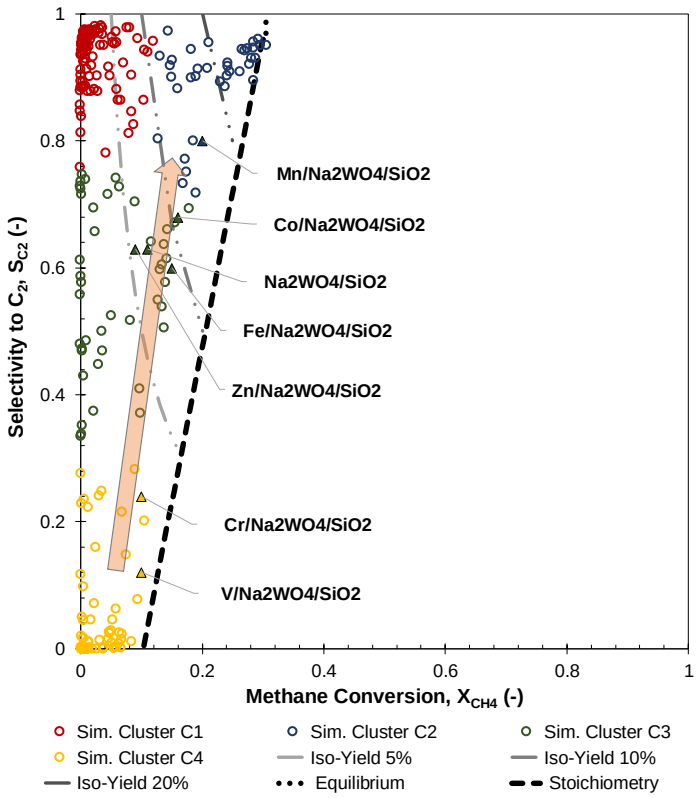


Figure 10. Four clusters of real (full triangles) and virtual (empty circles) catalysts, obtained with the k-means algorithm implemented in the Orange software[57] for the dataset of Malekzadeh et al.[81]. The colours distinguish the four clusters. Unique set of process conditions: T= 1048 K, p= 1 bar, $W/F_{CH4,0}$= 5.2 kg_{cat} s/$mol_{CH4,0}$, CH_4/O_2= 7.5. The arrow indicates the direction of growing electrical conductivity.

Step 5. The discriminating descriptors were considered to generate three *targeted* libraries, comprising C2, including the Mn oxide; C3, including the pure support and Zn, Fe and Co oxides; and C4, including Cr and V oxides. Figure 11 reports the distributions of these descriptors in the three *targeted* libraries.

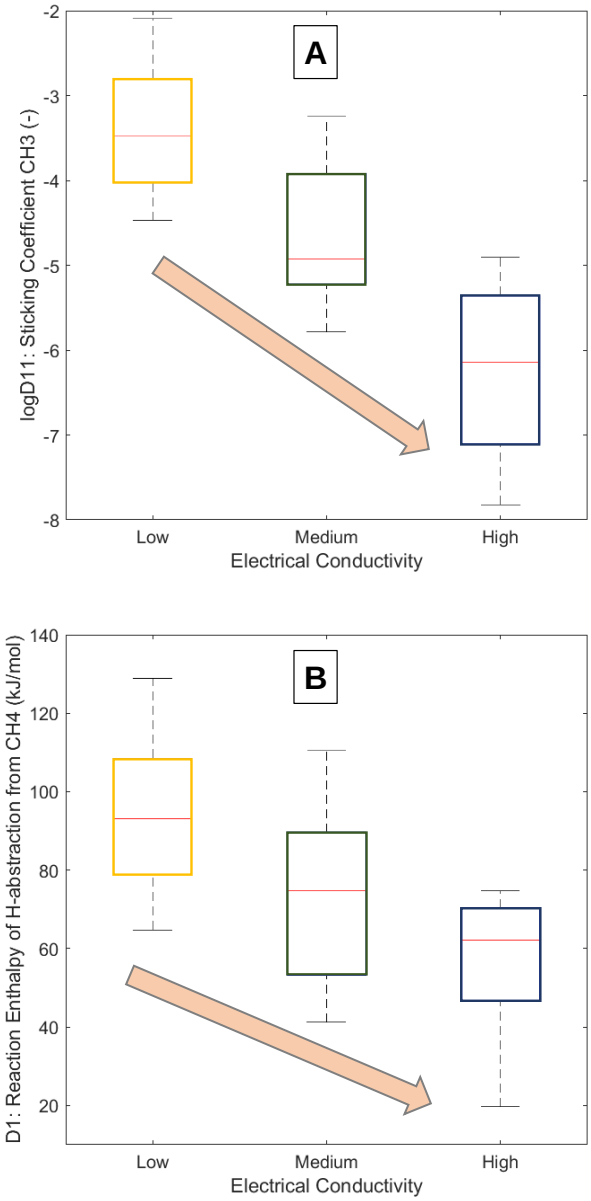


Figure 11. Comparison of the distributions of descriptors logD11: sticking coefficient of CH_3· (A) and D1: reaction enthalpy of H-abstraction from CH_4 (B) in the targeted libraries around C2, C3 and C4 (with decreasing electrical conductivity) from the dataset of Malekzadeh et al.[81].

It can be observed that both descriptors gradually follow the trend in performances: from low D1 and low D11 for the best performing *virtual* catalysts (C2) to high D1 and high D11 for the worst performing ones (C4).

Step 6. From properties of real catalysts to descriptor-property relationships The analysis of Figure 11 suggested that both D1 and D11 could be associated with the electronic properties of an OCM catalyst, and these two descriptors seemed to be correlated. A link between the reaction enthalpy of H-atom abstraction from CH_4 (D1) and the sticking coefficient of CH_3· (D11) on

595 one hand, and the electrical conductivity of a catalytic material on the other hand, was, hence,
596 identified.
597 The chemical background of these qualitative relationships is again related to the type of oxygen
598 species on the catalyst surface. In fact, the formation of basic lattice oxygen species (O^{2-}), which
599 in the previous section was associated with low D1, is favoured in highly conductive
600 materials[80]. Concerning D11, the different types of surface oxygen species are characterized by
601 a different degree of nucleophilicity, with the lattice oxygen (O^{2-}) being richer in electrons[39].
602 The formation of this oxygen species is favoured in p-type semiconductors[82], such as
603 $Mn/Na_2WO_4/SiO_2$ commonly known as one of the best performing OCM catalysts to date[83].
604 This oxygen species is also reported to be the most selective for OCM [84-86].
605

606 **4. Discussion**

607 The OCM case studies presented Section 3 demonstrate how the general methodology herein
608 proposed can be applied to a specific reaction. In the first case study, which was extremely
609 valuable in visualizing and interpreting the methodology, no statistically relevant descriptor-
610 composition relationship could be established, because the diversity in composition was not
611 reflected in a diversity of performances. A possible approach is to move from literature data to
612 own experimental data, in which compositions and process conditions are tuned to maximize
613 the spread in performance. As typical in kinetic studies, and relevant also for the approach
614 proposed in this work, experimental data obtained at low conversions and the evaluation of
615 selectivity as a function of increasing conversion can provide higher quality information rather
616 than single point evaluations obtained at high or complete conversion. Nevertheless, depending
617 on the investigated catalytic system, the relationship between composition and descriptors
618 might still not be evident, especially considering that an identical composition could result in
619 very different catalytic results depending on the synthesis technique. The quality of information

contained in the dataset is therefore of utmost importance. Other, more specific properties, such as strength of basic sites and electrical conductivity, are measured after synthesis and are therefore better suited for establishing descriptor-property relationships. The latter approach was followed for the two other case studies, suggesting that:

- Descriptor 1, i.e. the reaction enthalpy of H-abstraction from CH_4 by O*, could be linked to strength of the basic oxygen sites on the catalyst surface and to the electrical conductivity;
- Descriptor 2, i.e. the chemisorption enthalpy of O_2, could be linked to the strength of the basic oxygen sites on the catalyst surface;
- Descriptor 11, i.e. the sticking coefficient of CH_3 on O*, could be linked to the electrical conductivity.

Overall, a combination of medium-low enthalpy of H-abstraction from CH_4 (40 - 80 kJ/mol_{CH4}), medium-high oxygen chemisorption (200 - 280 kJ/mol_{O2}) and low sticking coefficient of CH_3· (10^{-8} - 10^{-5}) pointed to the presence of selective lattice oxygen (O^{2-}) as active species on the catalyst surface. With respect to different surface oxygen species, the OCM kinetic model only assumes a single type species O*. For that reason, in the past it was not possible to relate a descriptor specifically to it. In a previous work from our research group[32], performance optimization through kinetic model simulation led to ideal descriptor values in line with the ranges obtained in the present study (D1= 72.5 kJ/mol_{CH4}, D2= 224.1 kJ/mol_{O2}, D11= 10^{-7}). At the time, no further indication could be provided about the structure of an optimal catalyst. Thanks to the above descriptor-property relations, it is now possible to confirm that the electronic properties, affecting the nature of the surface oxygen species[82], should be the focus of the rational design of OCM catalysts. The desired relationship between properties and composition can be ultimately retrieved from specialized literature on tuning electrical conductivity and basicity via adequate doping of metal oxides[77, 80, 82, 87]. More specifically, basicity is a fundamental property of alkaline earth and rare-earth metal oxides and it can be

645 tuned via the addition of alkali (Li, Na, K, Rb, Cs) or alkaline-earth (Ca, Sr, Ba) metals[87].
646 Additionally, Alkaline earth oxides (CaO, SrO, BaO) and many rare earth oxides (Sm_2O_3,
647 La_2O_3) are p-type conductors[82] and both the oxygen-ion conductivity and p-type conductivity
648 can be increased by adding dopants with a lower valence compared to the host metal cation, to
649 create oxygen vacancies[80]. Whether a dopant is effectively incorporated in the host metal lattice,
650 however, depends on how compatible are the radii of the dopant and host metal cations[87].
651 With respect to the methodology itself, the OCM case study demonstrated its applicability to
652 small and diverse datasets which cannot be treated by other mathematical tools, such as machine
653 learning algorithms. This aspect was highlighted by Schmack et al.[49] as one of the major
654 limitations in the exploitation of available literature data in catalyst design. Diversity often
655 refers not only to composition, but also to process conditions at which *real* catalysts are tested.
656 Statistical approaches provide no guidelines in this respect. As an example[48], some OCM
657 catalysts have been predicted, via a random forest classification algorithm, to exceed the desired
658 30% C_2 yield, but were proposed to be operated at $CH_4/O_2 \leq 2$. However, this value, despite
659 being reported for laboratory-scale experiments, prospects an operating scenario which is most
660 probably unlikely in an industrial scale OCM process, due to the safety concerns associated
661 with it[88]. In a recent work[49], the operating temperature and the reactants ratio CH_4/O_2 were
662 included as variables in a multi-linear regression of the experimental C_2 yields. However, it was
663 mentioned that the analysis was not affected, neither qualitatively nor quantitatively, by the
664 presence of these two terms. Taking into account the high dependence of OCM catalytic
665 performance on these two process variables, this suggests that the impact of these process
666 variables was not fully accounted for by the proposed statistical model. The kinetic model at
667 the heart of the approach herein developed enables to overcome this limitation by explicitly
668 accounting for the process conditions case by case so to reproduce a wide variety of catalytic
669 performances and potentially optimize those.

It is also worth highlighting that the broad range of conversions and selectivities simulated in the present work were always within stoichiometric and thermodynamic boundaries. This was ensured by including thermodynamic consistency in the fundamental kinetic model, which is not guaranteed for a purely statistical model.

Given the systematic nature of the proposed approach, it is believed that it can be applied to other reactions provided that a kinetic model is at hand and experimental data, diverse in properties and performance, are available or can be generated. Some potential applications are dry methane reforming[89, 90] and biodiesel production from fatty acids transesterification[91, 92].

Being at proof of concept, the efficiency of the algorithm and uniformity of the statistical tools are not optimal yet as several software and development environments have been used. The integration of all methodological steps in a single software architecture could turn this methodology into an actual plug-and-play tool aimed at assisting both experimental and computational scientists in the design of new, better performing catalysts.

5. Conclusions

The methodology herein elaborated, represents a step forward in tackling one of the major challenges in kinetics-driven design of heterogeneous catalysts, namely the translation of a set of kinetically-relevant descriptors into tuneable properties of a catalytic material. A combination of fundamental kinetic modelling and carefully selected statistical tools, such as space-filling design of numerical experiments and k-means clustering, has been exploited to establish descriptor-property relationships. Important advantages of the developed approach are: a) the possibility to directly attribute chemical meaning to the results of the statistical tests, thanks to the deployment of detailed, thermodynamic consistent kinetic models, and b) the opportunity to start from existing literature data, even in case only small (< 10 samples) datasets are available.

Applied to OCM, it has been shown for the first time how chemically sound, qualitative relationships between catalyst descriptors, such as the H-abstraction enthalpy from CH_4 and chemisorption enthalpy of O_2, and measurable, tuneable properties of *real* catalytic materials, such as the strength of basic sites and electrical conductivity, were established. These relations confirmed that the electronic properties of the catalyst are related to lattice oxygen (O^{2-}) and are a key point of interest for the design of improved OCM catalysts.

In a broader perspective, given the conceptual and systematic approach, it is believed that the applicability of this approach can be extended to other reactions for which kinetics-driven catalyst design is relevant. Furthermore, the concepts and tools presented, such as kinetic simulations of libraries of *virtual* catalysts, can be applied to other fields than catalyst design only. For instance, the extremes in performance indicators of *targeted* catalyst libraries can serve as basis for techno-economic evaluations to explore the boundaries of chemical processes. Additionally, catalyst libraries targeted around realistic experimental data can also be exploited in the comparative simulations of alternative reactor configurations for which experiments are not available yet. This brings the proposed approach beyond the field of *in silico* catalyst design, and into the broader domain of computer-aided process design.

Supporting Information

S1. OCM kinetic model and catalyst descriptor summary; S2. Fast Flexible Filling (FFF) design of experiments and *discovery* library of *virtual* OCM catalysts; S3. additional information about dataset 1; S4. additional information about dataset 2; S5. additional information about dataset 3.

Glossary

Catalyst descriptor	Variable which impacts the kinetics of the reaction of interest and specifically depends on the type of catalyst (ex. chemisorption enthalpy, density of active sites, etc.).
Descriptors space	m-dimensional space, where m is the number of catalyst descriptors.
Design space	Hypercube in the *descriptors space*, which identifies the subset of the space determined by realistic value ranges for each of the m *catalyst descriptors*. The *design space* is the broadest possible during the first step of the first iteration of the methodology, aimed at generating a *discovery library*, and is progressively reduced during the *targeting* step.
Discriminating descriptor	*Catalyst descriptor* for which statistically significant differences (with 95% probability) are observed among different performance clusters, via Kruskal-Wallis/Mann-Whitney tests.
Virtual catalyst	Computer or 'in silico' representation of a catalytic material: each *virtual catalyst* is represented by a vector of m *catalyst descriptors*. Its performance can be simulated via a *numerical experiments* performed in a *numerical set-up.*
Real catalyst	Catalytic material which was synthesized, characterized and tested for the reaction of interest. In the present study, a *real catalyst* is represented by one or more numerical values representing its performance in the reaction of interest for a specific set of process conditions (ex. a conversion and selectivity, reported in literature) and by a property/composition which can be measured and represented again via a numerical value.
Discovery library	Collection of *virtual catalysts*, i.e. a $n \times m$ matrix of *catalyst descriptors*, in which each of the m columns identifies a *catalyst descriptor* and each of the n rows is associated to a *virtual catalyst*. It is obtained by sampling the broadest *design space* possible in the *descriptors space* via *FFF* during the first iteration of the methodology. The goal of the *discovery library* is to reproduce the full range of performances achievable for the reaction of interest.

Fast Flexible Filling (FFF)	Space-filling design of experiments method, based on hierarchical clustering. See section S2 of the Supporting Information.
Space-Filling Design	Group of design of experiments methodologies which sample both the external surface and the internal volume of the design hypercube, hence being well suited for deterministic computer experiments.
Targeting	Progressive reduction of the *design space*, by iteratively narrowing down the variability ranges of the *discriminating descriptors*, with the goal of generating *virtual catalysts* which are closer in performance to *real catalysts*.
Targeted library	Collection of *virtual catalysts*, i.e. a $n \times m$ matrix of *catalyst descriptors*, in which each of the m columns identifies a *catalyst descriptor* and each of the n rows is associated to a *virtual catalyst*. It is obtained by sampling the progressively smaller *design space* during the successive iterations of the methodology, after the *targeting* step. The goal of a *targeted library* is to reproduce a range of performances which are close the ones of *real catalysts.*
Numerical experiments	Reaction engineering simulations, aimed at obtaining the catalytic performances of *virtual catalysts* for the reaction of interest, in a set of process conditions which is the same reported for a dataset of *real catalysts* of interest. In these simulations, the values of the *catalyst descriptors* which define a certain *virtual catalyst* impact the kinetic parameters and, via changes in the reaction pathways, ultimately impact the performances.
Numerical set-up	Reactor + kinetic model which enables to simulate the catalytic performances of *virtual catalysts* for the reaction of interest, taking into account the variability in process conditions.

Nomenclature

C_a	cluster a, with $a \in [1, N_c]$
CH_4/O_2	reactants molar ratio at the inlet of the reactor
$D_{i,j}$	descriptor j for *virtual* catalyst i, with $j \in [1, m]$ and $i \in [1, n]$

DoE	Design of Experiments
Do(N)E	Design of (Numerical) Experiments
F_{CH4}	methane molar flowrate at the outlet of the reactor (mol/s)
$F_{CH4,0}$	methane molar flowrate at the inlet of the reactor (mol/s)
FFF	Fast Flexible Filling
l	number of performance indicators for the reaction of interest
m	number of catalyst descriptors in the kinetic model in exam
n	number of *virtual* catalysts
N_c	number of performance clusters
p	pressure (bar)
$P_{i,k}$	performance indicator k of catalyst i, with $k \in [1, l]$ and $i \in [1, n+q]$
q	number of *real* catalysts in the experimental dataset in exam
S_{C_2}	selectivity towards C_2 products (C_2H_6, C_2H_4, C_2H_2); $S_{C_2} = \frac{2\times(F_{C2} - F_{C2,0})}{F_{CH4,0} - F_{CH4}}$
T	temperature (K)
X_{CH_4}	methane conversion at the reactor outlet; $X_{CH_4} = \frac{F_{CH4,0} - F_{CH4}}{F_{CH4,0}}$
W	catalyst weight (kg)

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgments

The authors would like to thank Ana Obradović for the useful discussions in the initial stage of the work and David Vyncke for his feedback on the statistical methods. This work was supported by a research program agreement between TOTAL Research & Technology Feluy and Ghent University. The computational resources (Stevin Supercomputer Infrastructure) and

services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. JWT acknowledges the ERC for the PoC grant SERENiTi (GA n° 825783).

References

1. J. G. de Vries and S. D. Jackson, *Catal. Sci. Technol.*, 2012, **2**, 2009-2009.
2. S. Waclawek, V. V. T. Padil and M. Cernik, *Ecol. Chem. Eng. S*, 2018, **25**, 9-34.
3. J. K. Nørskov and T. Bligaard, *Angew. Chem. Int. Edit*, 2013, **52**, 776-777.
4. J. J. Bravo-Suárez, R. V. Chaudhari and B. Subramaniam, in *Novel Materials for Catalysis and Fuels Processing*, American Chemical Society, 2013, vol. 1132, ch. 1, pp. 3-68.
5. P. Cong, R. D. Doolen, Q. Fan, D. M. Giaquinta, S. Guan, E. W. McFarland, D. M. Poojary, K. Self, H. W. Turner and W. H. Weinberg, *Angew. Chem. Int. Edit.*, 1999, **38**, 483-488.
6. J. K. Nørskov, T. Bligaard, J. Rossmeisl and C. H. Christensen, *Nat. Chem.*, 2009, **1**, 37.
7. B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, 2311-2323.
8. T. Bligaard, R. M. Bullock, C. T. Campbell, J. G. Chen, B. C. Gates, R. J. Gorte, C. W. Jones, W. D. Jones, J. R. Kitchin and S. L. Scott, *ACS Catal.*, 2016, **6**, 2590-2602.
9. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73.
10. A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas and C. Phillips, *Scientific Data*, 2018, **5**, 180053.
11. J. M. Caruthers, J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, C. M. Snively, A. Bhan, S. Katare and G. Oskarsdottir, *J. Catal.*, 2003, **216**, 98-109.
12. K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T. N. Nguyen, S. Nishimura and T. Taniike, *ChemCatChem*, 2019, **11**, 1146-1152.
13. G. Rothenberg, *Catal. Today*, 2008, **137**, 2-10.
14. E. J. Ras, B. McKay and G. Rothenberg, *Top. Catal.*, 2010, **53**, 1202-1208.
15. E. J. Ras and G. Rothenberg, *RSC Adv.*, 2014, **4**, 5963-5974.
16. J. R. Kitchin, *Nat. Catal.*, 2018, **1**, 230-232.
17. A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders and R. Fushimi, *Acs Catal*, 2018, **8**, 7403-7429.
18. V. P. Ananikov, I. P. Beletskaya, G. G. Aleksandrov and I. L. Eremenko, *Organometallics*, 2003, **22**, 1414-1421.
19. G. Rostamikia and M. J. Janik, *Energy Environ. Sci.*, 2010, **3**, 1262-1274.
20. C. Liu, D. Shen, R. M. Sebastián, J. Marquet and R. Schönfeld, *Macromolecules*, 2011, **44**, 4616-4622.
21. K. Van der Borght, K. Toch, V. V. Galvita, J. W. Thybaut and G. B. Marin, *Catalysts*, 2015, **5**, 1948-1968.
22. S. Rangarajan, C. T. Maravelias and M. Mavrikakis, *J. Phys. Chem. C*, 2017, **121**, 25847-25863.

23. A. J. Medford, C. Shi, M. J. Hoffmann, A. C. Lausche, S. R. Fitzgibbon, T. Bligaard and J. K. Nørskov, *Catal. Lett.*, 2015, **145**, 794-807.
24. S. Katare, J. M. Caruthers, W. N. Delgass and V. Venkatasubramanian, *Ind. Eng. Chem. Res.*, 2004, **43**, 3484-3512.
25. K. Toch, J. W. Thybaut, M. A. Arribas, A. Martínez and G. B. Marin, *Chem. Eng. Sci.*, 2017, **173**, 49-59.
26. J. W. Thybaut and G. B. Marin, *J. Catal.*, 2013, **308**, 352-362.
27. I. R. Choudhury, K. Hayasaka, J. W. Thybaut, C. S. Laxmi Narasimhan, J. F. Denayer, J. A. Martens and G. B. Marin, *J. Catal.*, 2012, **290**, 165-176.
28. P. S. F. Mendes, F. M. Mota, J. M. Silva, M. F. Ribeiro, A. Daudin and C. Bouchy, *Catal. Sci. Technol.*, 2017, **7**, 1095-1107.
29. J. W. Thybaut, I. R. Choudhury, J. F. Denayer, G. V. Baron, P. A. Jacobs, J. A. Martens and G. B. Marin, *Top. Catal.*, 2009, **52**, 1251-1260.
30. S. R. Horton and M. T. Klein, *Energy Fuels*, 2014, **28**, 37-40.
31. M. Salciccioli and D. G. Vlachos, *ACS Catal*, 2011, **1**, 1246-1256.
32. J. W. Thybaut, J. J. Sun, L. Olivier, A. C. Van Veen, C. Mirodatos and G. B. Marin, *Catal. Today*, 2011, **159**, 29-36.
33. A. Zunger, *Nature*, 2019, **566**, 447-449.
34. J. K. Nørskov, F. Abild-Pedersen, F. Studt and T. Bligaard, *Proceedings of the National Academy of Sciences*, 2011, **108**, 937-943.
35. F. Studt, F. Abild-Pedersen, T. Bligaard, R. Z. Sørensen, C. H. Christensen and J. K. Nørskov, *Science*, 2008, **320**, 1320-1322.
36. F. Studt, I. Sharafutdinov, F. Abild-Pedersen, C. F. Elkjær, J. S. Hummelshøj, S. Dahl, I. Chorkendorff and J. K. Nørskov, *Nat. Chem.*, 2014, **6**, 320.
37. C. A. Wolcott, A. J. Medford, F. Studt and C. T. Campbell, *J. Catal.*, 2015, **330**, 197-207.
38. C. Stegelmann, A. Andreasen and C. T. Campbell, *J. Am.Chem. Soc.*, 2009, **131**, 8077-8082.
39. Y. Gambo, A. A. Jalil, S. Triwahyono and A. A. Abdulrasheed, *J. Ind. Eng. Chem.*, 2018, **59**, 218-229.
40. R. J. Kee, C. Karakaya and H. Y. Zhu, *Proc. Combust. Inst.*, 2017, **36**, 51-76.
41. M. Y. Sinev, Z. T. Fattakhova, V. I. Lomonosov and Y. A. Gordienko, *J. Nat. Gas Chem.*, 2009, **18**, 273-287.
42. J. A. Labinger, *Catal. Lett.*, 1988, **1**, 371-375.
43. E. V. Kondratenko, T. Peppel, D. Seeburg, V. A. Kondratenko, N. Kalevaru, A. Martin and S. Wohlrab, *Catal. Sci. Technol.*, 2017, **7**, 366-381.
44. L. Olivier, S. Haag, H. Pennemann, C. Hofmann, C. Mirodatos and A. C. van Veen, *Catal. Today*, 2008, **137**, 80-89.
45. E.C. Scher, F.R. Zurcher, J.M. Cizeron, W.P. Schammel, A. Tkachenko, J. Gamoras, D. Karshtedt and G. Nyce, *US Pat.*, 9,718,054, 2017.
46. U. Zavyalova, M. Holena, R. Schlogl and M. Baerns, *ChemCatChem*, 2011, **3**, 1935-1947.
47. E. V. Kondratenko, M. Schluter, M. Baerns, D. Linke and M. Holena, *Catal. Sci. Technol.*, 2015, **5**, 1668-1677.
48. K. Takahashi, I. Miyazato, S. Nishimura and J. Ohyama, *ChemCatChem*, 2018, **10**, 3223-3228.
49. R. Schmack, A. Friedrich, E. V. Kondratenko, J. Polte, A. Werwatz and R. Kraehnert, *Nat. Commun.*, 2019, **10**, 441.
50. V. I. Alexiadis, J. W. Thybaut, P. N. Kechagiopoulos, M. Chaar, A. C. van Veen, M. Muhler and G. B. Marin, *Appl. Catal. B-Environ.*, 2014, **150**, 496-505.

51. V. I. Alexiadis, M. Chaar, A. C. van Veen, M. Muhler, J. W. Thybaut and G. B. Marin, *Appl. Catal. B-Environ.*, 2016, **199**, 252-259.
52. D. Farrusseng, C. Klanner, L. Baumes, M. Lengliz, C. Mirodatos and F. Schuth, *QSAR Comb. Sci.*, 2005, **24**, 78-93.
53. G. Morra, D. Farrusseng, C. Bouchy and S. Morin, *Oil and Gas Science and Technology*, 2013, **68**, 487-504.
54. A. Trevino, https://www.datascience.com/blog/k-means-clustering, https://www.datascience.com/blog/k-means-clustering, (accessed 05/10/2018).
55. SAS, JMP 13.2.1, https://www.jmp.com).
56. P. N. Kechagiopoulos, J. W. Thybaut and G. B. Marin, *Ind. Eng. Chem. Res.*, 2014, **53**, 1825-1840.
57. J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik and B. Zupan, *J. Mach. Learn. Res.*, 2013, **14**, 2349-2353.
58. J. P. C. Kleijnen, S. M. Sanchez, T. W. Lucas and T. M. Cioppa, *INFORMS J. Comput.*, 2005, **17**, 263-289.
59. V. C. P. Chen, K.-L. Tsui, R. R. Barton and M. Meckesheimer, *IIE Transactions*, 2006, **38**, 273-291.
60. S. S. Garud, I. A. Karimi and M. Kraft, *Comput. Chem. Eng.*, 2017, **106**, 71-95.
61. L. Pronzato and W. G. Müller, *Stat. Comput.*, 2012, **22**, 681-701.
62. R. Lekivetz and B. Jones, *Qual. Reliab. Eng. Int.*, 2015, **31**, 829-837.
63. J. L. Loeppky, J. Sacks and W. J. Welch, *Technometrics*, 2009, **51**, 366-376.
64. A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding and C.-T. Lin, *Neurocomputing*, 2017, **267**, 664-681.
65. P. J. Rousseeuw, *J. Comput. Appl. Math.*, 1987, **20**, 53-65.
66. D. Arthur and S. Vassilvitskii, New Orleans, Louisiana, 2007.
67. S. Zacks, in *Parametric Statistical Inference*, ed. S. Zacks, Pergamon, 1981, DOI: https://doi.org/10.1016/B978-0-08-026468-4.50007-8, pp. 1-14.
68. P. K. Sen, *J. Nonparametr. Stat.*, 1991, **1**, 1-32.
69. W. H. Kruskal and W. A. Wallis, *J. Am. Stat. Assoc.*, 1952, **47**, 583-621.
70. H. B. Mann and D. R. Whitney, *Ann. Math. Stat.*, 1947, **18**, 50-60.
71. A. M. Mood, *Ann. Math. Stat.*, 1954, **25**, 514-522.
72. Statistics How To, http://www.statisticshowto.com/).
73. J. Panerati, D. Sciuto and G. Beltrame, in *Handbook of Hardware/Software Codesign*, Springer Science+Business Media Dordrecht, 2017.
74. K. Ito, I. Couckuyt, R. d'Ippolito and T. Dhaene, *Appl. Soft Comput.*, 2016, **43**, 337-346.
75. L. Pirro, A. Obradović, B. D. Vandegehuchte, G. B. Marin and J. W. Thybaut, *Ind. Eng. Chem. Res.*, 2018, **57**, 16295-16307.
76. S. Kuś, M. Otremba and M. Taniewski, *Fuel*, 2003, **82**, 1331-1338.
77. V. H. Rane, S. T. Chaudhari and V. R. Choudhary, *J. Nat. Gas Chem.*, 2008, **17**, 313-320.
78. A. M. Maitra, I. Campbell and R. J. Tyler, *Appl. Catal. A-General*, 1992, **85**, 27-46.
79. A. A. Davydov, *Chem. Eng. Technol.*, 1995, **18**, 7-11.
80. H. Borchert and M. Baerns, *J. Catal.*, 1997, **168**, 315-320.
81. A. Malekzadeh, A. Khodadadi, M. Abedini, M. Amini, A. Bahramian and A. K. Dalai, *Catal. Commun.*, 2001, **2**, 241-247.
82. Z. Zhang, X. E. Verykios and M. Baerns, *Cataly. Rev.*, 1994, **36**, 507-556.
83. S. Arndt, T. Otremba, U. Simon, M. Yildiz, H. Schubert and R. Schomäcker, *Appl. Catal. A-General*, 2012, **425-426**, 53-61.

84. Y. Gordienko, T. Usmanov, V. Bychkov, V. Lomonosov, Z. Fattakhova, Y. Tulenin, D. Shashkin and M. Sinev, *Catal. Today*, 2016, **278**, 127-134.
85. V. Fleischer, R. Steuer, S. Parishan and R. Schomäcker, *J. Catal.*, 2016, **341**, 91-103.
86. I. Kim, G. Lee, H. B. Na, J.-M. Ha and J. C. Jung, *Molecular Catalysis*, 2017, **435**, 13-23.
87. D. Filkova, D. Wolf, G. Gayko, M. Baerns and L. Petrov, *Appl. Catal. A-General*, 1997, **159**, 33-44.
88. S. Stunkel, H. Trivedi, H. R. Godini, S. Jaso, N. Holst, S. Arndt, J. Steinbach and R. Schomacker, *Chem. Ing. Tech.*, 2012, **84**, 1989-1996.
89. S. Praserthdam and P. B. Balbuena, *Catal. Today*, 2018, **312**, 23-34.
90. A. N. Şener, M. E. Günay, A. Leba and R. Yıldırım, *Catal. Today*, 2018, **299**, 289-302.
91. N. Alper Tapan, R. Yıldırım and M. Erdem Günay, *Biofuel. Bioprod. Bior.*, 2016, **10**, 422-434.
92. F. Trejo-Zárraga, F. de Jesús Hernández-Loyo, J. C. Chavarría-Hernández and R. Sotelo-Boyás, in *Biofuels*, ed. K. Biernat, IntechOpen, 2018, ch. 8.